National Aeronautics and Space Administration

EXPLORE SCIENCE

Workshop on Maximizing the Scientific Return of NASA Data

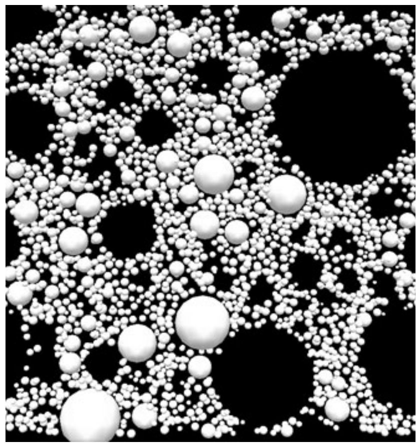
October 30-31, 2018

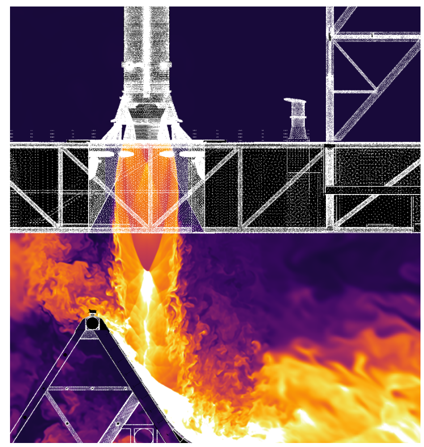
www.nasa.gov

All studies, findings and recommendations in these deliverables have been submitted to the Strategic Data Management Working Group (SDMWG) and to officials at NASA HQ. The opinions expressed in these materials do not reflect NASA’s concurrence, approval, or indicate steps to implementation.

Executive Summary

The *Workshop on Maximizing the Scientific Return of NASA Data* was convened by the Science Mission Directorate (SMD) in Washington, D.C. on October 30-31, 2018. The workshop brought together thought leaders from NASA, academia, industry and government and international partners to gather community input on a new, Directorate-wide approach to leveraging NASA research data to advance informational technology that enables groundbreaking scientific research. The volume of data in NASA's science archives is expected to double over the next five years and grow exponentially thereafter as new Missions are launched, presenting unique scientific opportunities as well as significant challenges for data management, access and analysis. Historically, NASA's management of data and computing resources has been conducted on a Division or Mission basis, with limited consideration for enabling interdisciplinary research.





SMD leads the scientific community in the implementation of effective methods to collect, manage and distribute scientific data. SMD's next milestone in pioneering data innovation is through the development of a new Strategic Plan for Scientific Data and Computing (Strategic Plan). This new Strategic Plan will guide the evolution of the array of data and computing systems supporting research across the four science areas within SMD over the next five years. The focus of the two-day interactive workshop was to identify and discuss:

- Best approaches for improving data storage and data processing capabilities to scale for increased demand and volumes of data
- Best practices to improve discoverability and ease of use for data
- Sustainable methods to expand SMD efforts to maintain free/easily accessible databases
- Data processing applications across disciplines and among different agencies (internal, external, domestic and global partners)
- How to manage the potential benefits and pitfalls of big data and computing

Workshop participants included over 150 individuals from diverse arenas of science and technology, including subject matter experts (SMEs) in data science, computing, archives and management.

Participants made several actionable recommendations regarding how NASA should approach data management to accommodate the speed of technology innovation and stakeholder community evolution. Participants agreed that successful data management plans must incorporate open science principles, increased educational opportunities relevant to data science and cloud-based solutions.

Workshop on Maximizing the Scientific Return of NASA Data – Overview

Purpose and Goals

NASA SMD's Strategic Data Management Working Group (SDMWG) convened a workshop on October 30-31, 2018 in Washington, D.C. The workshop focused on engaging with members of the scientific community to discuss science data systems, including high-end computing, to promote more efficient and effective data management across SMD divisions and enable cross-disciplinary discovery and analysis of science data. The workshop aimed to serve as a vehicle for sharing information, ideas and lessons learned to articulate realistic recommendations for the Strategic Plan.

The development of a new SMD-wide data management strategy will align the advances in information technology with the unique needs of science data systems and computing. This union lets the Strategic Plan both inform technology investments and provide a roadmap for how SMD can partner with other organizations, within NASA and externally, to enable greater scientific discovery.

The objective of the Strategic Plan is to articulate a strategy that has four overall goals:

1. Improve discovery and access of all SMD data for immediate benefit to science data users and the overall user experience.
2. Identify researchers and use cases that are both large-scale and cross-disciplinary to inform future science data system capabilities.
3. Champion robust theory programs that are firmly based on NASA's observations.
4. Modernize science data and computing systems to improve efficiency and enable new technology and analysis techniques for scientific discovery and commercial use.

Description of Workshop

Over 150 participants attended the two-day workshop in-person or virtually. Participants and presenters represented the science community, NASA domestic and international partners, leaders in the IT industry and small businesses. They included scientists, researchers, engineers, archivists, educators, entrepreneurs and post-doc students. Figure 1 below shows a breakdown of workshop participants by their organizational affiliation.

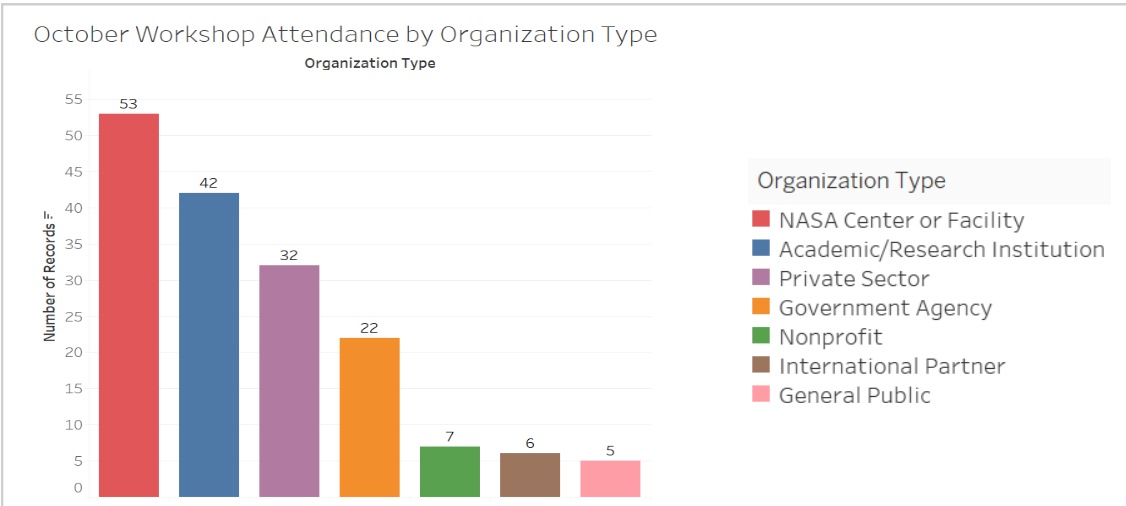


Figure 1: October Workshop Attendance by Organization Type

The workshop consisted of eight technical sessions, a briefing on NASA cloud computing and a lightning talk session. The co-chairs of the SDMWG kicked off the workshop with a talk on the inspiration for the event and the core principles of SMD that brought the work group together.

The workshop was designed to engage experts specializing in NASA research, management of big data best practices and information technology to:

- Share and identify best practices
- Identify cross-cutting challenges and opportunities to enable groundbreaking science and increase use of NASA data over the next five years
- Identify and develop successful relationships with other agencies, industry and academic partners

Summary of Workshop Discussions

There were nine workshop sessions that featured presentations by small groups of SMEs, followed by a panel and Q&A session focused on the following subjects:

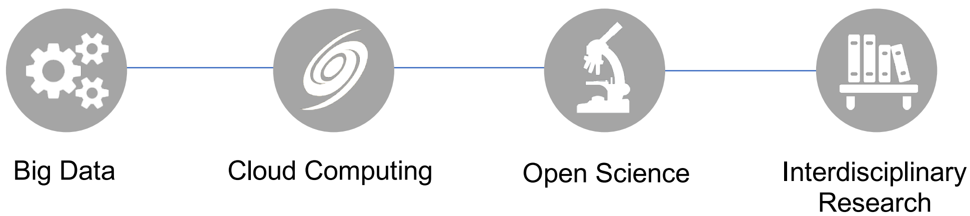
- The current state of NASA's data and computing resources
- How to continue to enable scientific research and collaborations
- Current and potential challenges for analyzing scientific data

- Best practices and lessons learned concerning data management policies and stewardship
- Where data curation is moving in the future

The workshops highlighted unique data management challenges across science disciplines. Participants developed actionable recommendations as major takeaways for all to analyze.

Key Themes

Workshop discussions encompassed four overall themes:



The following sections summarize the workshop discussions in the context of these themes. For more information on each session, see the Appendix.

Big Data

The rapid growth in the amount of scientific data collected from NASA Missions in recent years is enormous. NASA supports hundreds of thousands of collections of observational data, model data and experimental data. Participants in the workshop identified a variety of data management challenges including but not limited to: data discovery, data management, analysis workflow and analysis and infrastructure. However, the workshop speakers highlighted that the true challenge inherent in the data is its complexity, not its volume. Many speakers acknowledged that the definition of "complexity" varied amongst the science disciplines. Multiple definitions include the variety of observed data sources such as observation techniques, data types and data analysis methods/tools. In addition to the explosion of data, new data science and machine learning techniques have facilitated new ways to analyze data – and revealed new challenges to managing the data.

One of the areas of concern discussed was setting standards for, and the preservation of, metadata to facilitate usability for research data. Metadata preservation strategies must consider a long-term view of the value of data, preparing for decades or even centuries in the future. Suggestions included generating metadata automatically, adding standardized metadata requirements to contracts, using new techniques (e.g., machine learning) to apply metadata and appending metadata to the results of data analysis – not just to the raw data.

Cloud Computing

Cloud computing was a recurring theme in most of the presentations throughout the workshop. Cloud computing offers benefits such as the ability to analyze data at scale, analyze multiple data sets together easily and avoid lengthy expensive moves of large data sets, allowing scientists to work on data “in place.” NASA was a pioneer in cloud computing, having established its own community cloud computing data center called Nebula in at the Ames Research Center (ARC) in 2009. However, most of NASA’s IT assets are still acquired and managed according to the traditional infrastructure model. Cloud computing is a significant departure from that model. Utilizing it effectively will require technical training as well as a shift in how NASA teams and programs think about, budget for, and procure IT resources. Using cloud computing also requires different workflows for processing, storing and accessing data.

The workshop participants acknowledged there are several types of cloud deployment and no single type of cloud computing solution is right for every enterprise. Most speakers were in favor of NASA continuing to leverage this technology across SMD and the other Mission Directorates. Speakers put forth three main arguments in favor of cloud computing: reduced pricing, reduced startup time and the relative ease of conducting analytics in the cloud.

Using the cloud for analytics is less expensive than maintaining an in-house computing capability, since users are paying only for resources used and not for maintaining under-utilized computing and network capacity. However, this economic advantage only applies to computation in the cloud, not storage.

Another key advantage to cloud computing is that it provides “quick start” capabilities that allow users to configure, initiate, run and decommission instances as needed. Users do not have to build their own tools, since suites of tools are already provided as part of the service. Nor do they have to be concerned about system administration or hardware maintenance – the cloud provider is responsible for that.

Additionally, participants advocated cloud computing as an essential enabling technology for performing seamless server-side analytics on large collections of scientific data.
Speakers noted a few downsides to cloud computing as well, including the cost of data access. NASA is required in many instances to provide access to its data for free, but in a commercial cloud environment it is challenging to limit costs while allowing unlimited downloads. Workshop participants also discussed the problem of moving data from one cloud to another and the importance of avoiding “vendor lock-in,” i.e., dependence on a single provider’s proprietary tools and/or data formats.

Open Science

The National Academies of Sciences “Open Source Software Policy Options for NASA Earth and Space Sciences” presentation focused on how NASA and the science community should approach and leverage open-source tools. Participants also discussed the benefits and pitfalls of open science. “Open science” is defined as making scientific research (including publications, data, physical samples and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional. Open science is transparent and accessible knowledge that is shared and developed through collaborative networks.

Discussion regarding the importance of and challenges inherent in open science were a constant throughout the workshop, covering topics such as providing unlimited access to data, source code and scientific methods. There was widespread agreement that increasing accessibility would produce dramatic improvements for NASA's scientific endeavors and the researchers and communities that use NASA data.

However, several speakers mentioned that incentives are required to encourage open science. For example, a researcher using their own modeling code can generate high-impact papers to get tenure, grants, etc. Under the current structure, releasing that source code and allowing others to use and build upon it would not allow a researcher any advantage, so incentivizing openness is key to obtaining user buy-in.

Workshop participants also discussed the importance of education, legal complications and ethics. One presentation addressed the legal complications that arise when open code is written by government employees. Finally, a few speakers, particularly those in health care fields, discussed how in some cases data cannot be open due to patient privacy, security or other legal issues. In summary, it was apparent that there are many misconceptions around open science and its potential implications, but that there was a large consensus that NASA should pursue support of open science in spite of the associated challenges.

Interdisciplinary Research

Science efforts are steadily becoming more interdisciplinary, which creates significant challenges for NASA to address. NASA should expect users from outside of a dataset's "principal discipline" to want access to Agency data and the Agency is strongly advised to consider meeting those users' expectations for accessibility when making datasets public and designing data products.

Several speakers also mentioned best practices for breaking down data silos and producing interdisciplinary research, such as using shared data models and indexing metadata to make it easily discoverable. As one of the workshop speakers stated, "Science of the future will require interdisciplinary collaboration across the SMD's science divisions and beyond as the specialties and science advance."

NASA is advised to consider how to accommodate and facilitate interdisciplinary efforts as it examines its funding and grant systems. In many academic spaces, multidisciplinary teams of individuals with different backgrounds, skillsets, and subject area knowledge are becoming more common. Currently, it is difficult for multidisciplinary research projects to secure grant funding, simply because NASA's grant systems assume single-discipline studies. Many participants also perceived a need for NASA to support a common "language" for interdisciplinary studies, since field-specific jargon can hinder collaboration.

Key Challenges

As NASA tries to adapt to the increasing data size and complexity of data, developing an effective policy will be necessary to overcome some key challenges. These challenges include misaligned incentives, lack of knowledge/skills in a rapidly changing field of science, a lack of infrastructure and difficulty in identifying trusted developers.

The first two challenges are related to people. Currently, incentives are lacking for researchers—both within NASA and in the broader research community—to engage in best practices of openness and collaboration. Furthermore, while the nature of science and scientific data is changing in the era of “Big Data,” most researchers lack the background and/or resources/funding in computer science, statistics, data science and machine learning to be able to make maximum use of the data.

The last two challenges relate to tools and infrastructure. Contemporary scientific data analysis requires the ability to perform seamless server-side analytics, which involves hosting data in a cloud environment (either commercial or NASA-owned). NASA needs to carefully evaluate and select which, if any, commercial cloud provider to use and must also avoid being “locked in” to a single cloud provider’s proprietary software and/or data formats.

Recommended Action Items for NASA

Workshop participants suggested that NASA implement an integrated policy and strategic approach to meeting data science and computing needs throughout SMD. Currently, solutions for meeting these needs are siloed across the divisions and do not facilitate interdisciplinary exchange or data sharing.

Any approach the Agency takes to better facilitating data sharing must address both the technological challenges and the people challenges inherent in managing and providing access to its data. One way to do this is to develop a standardized data management strategy and open science requirements to be integrated into future ROSES proposal language and grants. One of the recommendations that emerged from the workshop was to require researchers to cite not only publications in their proposals and reports, but also relevant datasets and availability or accessibility of research models.

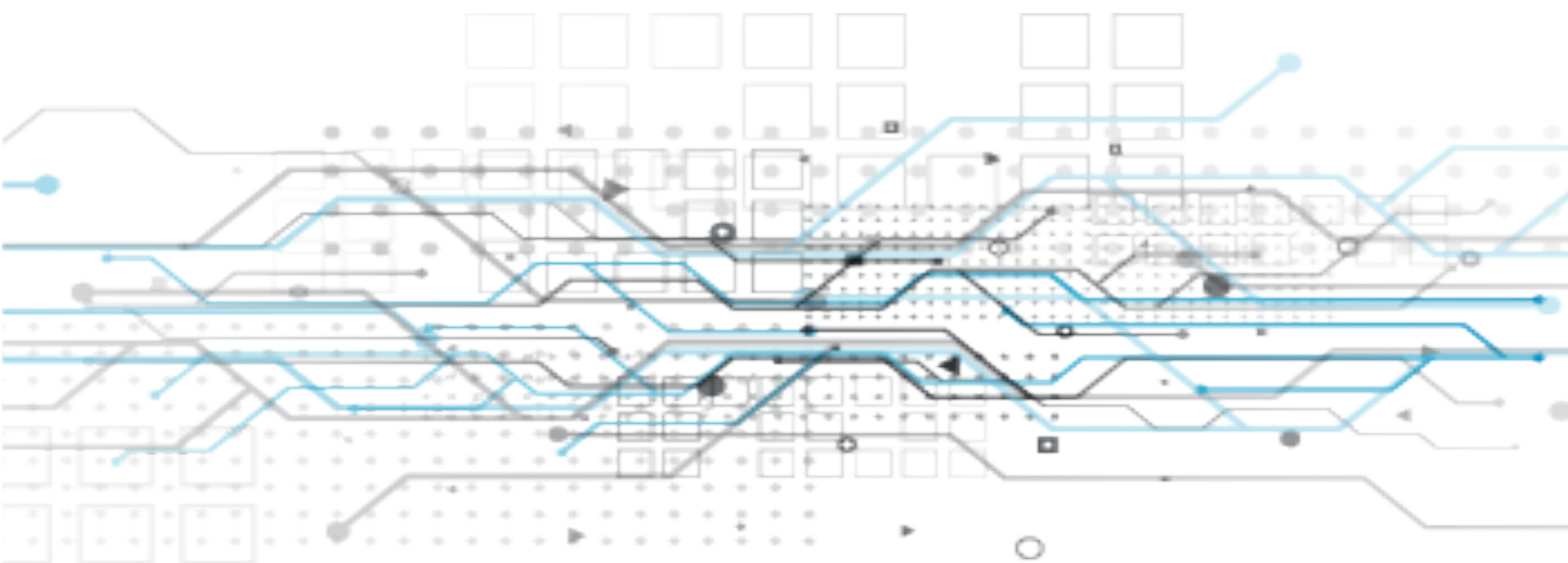
Workshop participants also recommended a data science training and awareness program (e.g. machine learning, artificial intelligence, and big data) for NASA employees and research partners. To remain at the cutting edge, domain scientists within NASA must have the opportunity and accessibility to learn new technologies and leverage the commercial sector’s vast capabilities to fill the gaps. The Agency can also participate and/or otherwise encourage collaborative research and open science by:

- Implementing a more comprehensive data management and computing policy that spans across all SMD divisions and can coordinate efforts among divisions
- Improving education around enhanced science applications and new techniques, in particular for domain scientists
- Leveraging in-house technologies (e.g. Nebula)

NASA may also choose to explore leveraging commercial cloud solutions for analytics and open data. When selecting and acquiring cloud services, the Agency can ensure that contract awards do not result in vendor lock-in by requiring easy transfer of data and algorithms between different clouds in cloud service contracts. The Agency is advised to weigh all aspects of cloud service pricing, including that used for storage versus computing and data access and storage fees. NASA can also consider leveraging commercial cloud services to supplement the Agency's own considerable investment in HPC.

Conclusion and Next Steps

The information gathered from the *Maximizing the Scientific Return of NASA's Data* workshop will play a vital role in the development of SMD's data management and computing strategy. By bringing together a diverse group of thought leaders from academia, industry and government, SMD was able to make connections and better understand shared challenges involved in managing big data. The overall outcome of the workshop indicates that while there are significant challenges to address, sharing lessons learned and best practices will foster a culture that facilitates working across multiple disciplines, advances science and enables more people to use scientific data. By implementing policies that actively promote data sharing, SMD can continue to meet the evolving needs of the science community.



Appendix: Overview of Workshop Sessions

SMD Introduction

- Ellen Gertsen, NASA Headquarters
- Kevin Murphy, NASA Headquarters

National Academy of Sciences Report on Open Code

- Chelle Gentemann, Earth and Space Research
- Mark Parson, Rensselaer Polytechnic Institute

State of Research – Science Questions

- Larry Di Girolamo, University of Illinois
- Jeff Kruk, Goddard Space Flight Center, NASA
- Tamas Gombosi, University of Michigan

State of Research – Art of the Possible

- Ian Foster, Argonne National Lab
- Amitava Bhattacharjee, Princeton University
- Daniela Huppenkothen, Data Intensive Research in Astrophysics and Cosmology (DIRAC) Institute
- Ralph McNutt, John Hopkins University Applied Physics Laboratory

State of Research – Training

- Tyler Erickson, Google Earth Outreach
- Jennifer Houchins, Shodor/XSEDE
- James Drake, University of Maryland
- Kelle Cruz, AstroPy
- Tracy Teal, Software Carpentry

Interagency Partners

- Ed Kearns, National Oceanic and Atmospheric Administration
- Tod Dabolt, Department of the Interior
- Manish Parashar, National Science Foundation
- Tom McGlynn, Goddard Space Flight Center, NASA
- Dina Paltoo, National Institute of Health

International Partners

- Shin-ichi Sobue (remote), Japan Aerospace Exploration Agency
- Paul Counet (remote), European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT)
- Yves Buhler (remote), EUMETSAT
- Nicolaus Hanowski , European Space Agency
- Jessica Severin, RIKEN
- Giuseppina Fabbiano, International Virtual Observatory Alliance (IVOA)

Cloud Computing at NASA

- Karen Petraska, NASA Headquarters

Industry Perspective

- Brett McMillen, Amazon Web Services
- Eric Pennaz, Google
- Hendrik Hamann, IBM T.J. Watson Research Center
- Alison Lowndes, NVIDIA
- Susie Adams, Microsoft

Data and Science Innovators

- Ian Schuler, DevSeed
- Daniel Crichton, NASA, JPL
- W. Kent Tobiska, Space Environment Technologies


NASA
EXPLORE
with us